DOCUMENT RESUME

ED 351 390 TM 019 228

AUTHOR	Van Velsor, Ellen; Leslie, Jean Brittain
TITLE	Feedback to Managers, Volume I: A Guide to Evaluating Multi-Rater Feedback Instruments.
INSTITUTION	Center for Creative Leadership, Greensboro, N.C.
REPORT NO	CCL-R-149; ISBN-0-912879-65-3
PUB DATE	Nov 91
NOTE	71p.; For Volume II, see TM 019 229.
AVAILABLE FROM	Center for Creative Leadership, 5000 Laurinda Drive, P.O. Box 26300, Greensboro, NC 27438-6300 ($15; $12 per item for orders of 2-49; $9 per item for orders of 50+).
PUB TYPE	Guides - Non-Classroom Use (055) -- Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Administrators; Behavior Patterns; *Evaluation Methods; Evaluators; *Feedback; Guidelines; Leadership Qualities; *Leadership Styles; *Managerial Occupations
IDENTIFIERS	Self Report Measures

ABSTRACT

"Feedback to Managers" is a two-volume report. Volume 1 presents a simple but comprehensive process for evaluating instruments designed to compare self-views to the views of others on multiple domains of managerial behavior. Sixteen steps in evaluating a multiple-perspective feedback questionnaire intended for management development are outlined sequentially, although they are not of equal importance. The most critical decisions come early in the process to enable the elimination of instruments that clearly do not meet the evaluator's needs. The following are the 16 steps: (1) find out what is available; (2) collect a complete set of materials; (3) compare your intended use to instrument characteristics; (4) examine the feedback scales; (5) familiarize yourself with the instrument-development process; (6) learn how items and feedback scales were developed; (7) find out how consistent scores tend to be; (8) assess basic aspects of validity (determine whether or not the instrument measures what it claims to measure); (9) think about face validity; (10) examine the response scale; (11) evaluate the feedback display; (12) understand how breakout of rater responses is handled; (13) learn what strategies are used to facilitate interpretation of scores; (14) look for development and support materials; (15) compare cost in terms of value for the price; and (16) consider length a minor issue. A 16-item checklist of the steps is included, as well as a glossary of technical terms. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN **OTHER THAN PAPER COPY** HAS BEEN GRANTED BY

MARCIA HOROWITZ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED351390

TM019228

FEEDBACK TO MANAGERS
Volume I:
A Guide to Evaluating Multi-rater Feedback Instruments

Ellen Van Velsor
Center for Creative Leadership

Jean Brittain Leslie
Center for Creative Leadership

PUTTING IDEAS INTO ACTION

A Center for Creative Leadership Applications Report

BEST COPY AVAILABLE

EDITORIAL POLICY

The Center for Creative Leadership publishes reports on the management and development of the human resource in complex organizations. Areas of special focus include leadership, executive and managerial development, individual and organizational creativity, and human resource systems. Primarily an outlet for the work of Center staff, the reports include theoretical, empirical, and practical articles whose purpose is to make ideas and knowledge from research accessible to managers and human resource specialists.

Thank you . . .

. . . for supporting the work of the Center through your purchase of this report. We hope you find it valuable. If you have comments, suggestions, or questions about our publications, please write to Bill Drath, Director of Publication, at the address below.

A Center for Creative Leadership Publication

©1991 by Center for Creative Leadership
Post Office Box 26300
Greensboro, North Carolina 27438-6300

(919) 288-7210

FEEDBACK TO MANAGERS
Volume I:
A Guide to Evaluating Multi-rater Feedback Instruments

Ellen Van Velsor
Center for Creative Leadership

Jean Brittain Leslie
Center for Creative Leadership

November 1991

©1991 by Center for Creative Leadership
5000 Laurinda Drive
Post Office Box 26300
Greensboro, North Carolina 27438-6300

Report Number 149
All Rights Reserved
ISBN 0-912879-65-3

About the Authors

Ellen Van Velsor is Director, Product Development Research, at the Center. She manages projects which focus on individual development as a result of feedback as well as those which focus on the development and evaluation of assessment technologies.

In addition to numerous articles published in the areas of personality assessment and transfer of training, Van Velsor has published extensively about women's career development. She is coauthor of *Breaking the Glass Ceiling: Can Women Reach the Top of America's Largest Corporations?* (1987, Addison-Wesley).

Van Velsor has a Ph.D. in sociology from the University of Florida.

Jean Brittain Leslie is a Research Analyst at the Center. Her current research focuses on psychometric analysis of Center instruments and cross-cultural applicability of leadership models.

Leslie has taught sociology at Elon College, Guilford Technical Community College, and Greensboro College, all in North Carolina.

She has an M.A. in sociology from the University of North Carolina at Greensboro.

Contents

PREFACE vi

ACKNOWLEDGEMENTS viii

INTRODUCTION 1

STEPS IN THE EVALUATION PROCESS 2

GETTING STARTED 2

STEP 1: FIND OUT WHAT IS AVAILABLE 2

STEP 2: COLLECT A COMPLETE SET OF MATERIALS 3

STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS 4

STEP 4: EXAMINE THE FEEDBACK SCALES 6

STEP 5: FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS 7

STEP 6: LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED 8

STEP 7: FIND OUT HOW CONSISTENT SCORES TEND TO BE 11

Homogeneity Within Scales: Basic Issues in Internal Consistency 12

Agreement Within Rater Groups: Basic Issues in Interrater Reliability 13

Additional Considerations 13

Stability Over Time: Basic Issues in Test-retest Reliability 14

Additional Considerations 15

STEP 8: ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE? 16

Additional Considerations: Ways of Defining Validity 20

Content Validity 20

Construct Validity 21

Concurrent and Predictive Validity 22

Criterion Issue 22

Reliability of Criterion Measure 24

There is No Perfect Instrument: Balance is the Key 24

CONSIDER POTENTIAL IMPACT 25

STEP 9: THINK ABOUT FACE VALIDITY 25

STEP 10: EXAMINE THE RESPONSE SCALE 26

STEP 11: EVALUATE THE FEEDBACK DISPLAY 28
Type of Display 28
Clarity of Feedback Display 29

STEP 12: UNDERSTAND HOW BREAKOUT OF RATER RESPONSES IS HANDLED 29

STEP 13: LEARN WHAT STRATEGIES ARE USED TO FACILITATE INTERPRETATION OF SCORES 31
Comparison to Norms 32
Highlighting Largest Self/Rater Differences 33
Item-level Feedback 33
Highlighting High/Low Scores 33
Comparison to Ideal 34
Importance to Job or Success 34
"Do More/Do Less" Feedback 35
Linking Strategies Together: Issues in Considering Impact of Feedback 35

STEP 14: LOOK FOR DEVELOPMENT AND SUPPORT MATERIALS 36
Support for the Manager's Development 36
Trainer Support 37
Scoring Process 38
Manual Scoring 38
Computer Scoring 39
Vendor Scoring 39
Computer Software 40

OTHER CONSIDERATIONS 41

STEP 15: COMPARE COST: VALUE FOR THE PRICE 41

STEP 16: CONSIDER LENGTH A MINOR ISSUE 42

CONCLUSION 43

REFERENCES 44

GLOSSARY OF TERMS 45

INSTRUMENT EVALUATION CHECKLIST 49

PREFACE

Feedback is a rare commodity in day-to-day organizational life, but it is a key to managerial effectiveness. One increasingly popular vehicle for getting feedback from one's boss, peers, and subordinates is the multiple-perspective, or "360-degree," feedback instrument. Taken as part of a management-development course or used as a stand-alone device, this kind of tool can enhance self-confidence by highlighting individual strengths and can facilitate greater self-awareness by pointing out areas in need of further development.

More than a decade ago the Center for Creative Leadership published *Feedback to Managers: A Comprehensive Review of Twenty-four Instruments* (Morrison, McCall, & DeVries, 1978). A popular publication among a long line of Center technical reports, it covered virtually all instruments available at the time that provided multiple-perspective feedback to managers.

Yet the field has grown tremendously since 1978, and continues to expand today, giving potential users of these products more choices than ever before. Although revision of the original report is long overdue, the reality today is that any one catalog of feedback instruments will be quickly obsolete. More than ever before, human-resource professionals need to have the ability to conduct thorough instrument evaluations if they are to keep up with the latest products available and get the one that best meets their needs.

This new version of *Feedback to Managers* is designed to assist you in that task. It is a two-volume report, presenting, in volume I, a simple but comprehensive process for evaluating instruments designed to compare self-view to the views of others on multiple domains of managerial behavior. This model is followed, in volume II, by a comparison of sixteen of the better feedback instruments we were able to obtain. The reader can use this volume to gain an appreciation of what some of the most experienced instrument developers have been able to achieve in terms of reliability and validity and in terms of constructing high-impact feedback displays. Complete information on these sixteen is presented in the Summaries section of volume II.

It is our hope that this approach will be of enduring benefit, in that managers will be able to use these guides to get an overview

of the field today and to learn a process by which to evaluate any feedback instrument that comes across their desks today or in the foreseeable future.

ACKNOWLEDGEMENTS

We would like to express our appreciation to the many reviewers whose comments, ideas, and criticisms greatly improved the quality of this manuscript. These include David DeVries, Janet Spence, Russ Moxley, Michael Hopp, Maxine Dalton, Lloyd Bond, and members of the Center's Writers' Group. We are also indebted to Clark Wilson, Walt Tornow, and David Campbell for several ideas presented here.

INTRODUCTION

This volume, part one of a two-volume set, presents a nontechnical, step-by-step process you can use to evaluate any multiple-perspective feedback questionnaire intended for management development. Although we have simplified this process as much as possible, it still will require some effort on your part—but effort that will pay off in terms of your having a high-quality instrument that best meets your needs. You can read and use the present volume without reading volume II.

The steps in evaluating an instrument are laid out here sequentially. Yet all steps are not equal in complexity or importance. We suggest that you make the most critical decisions early in the process; in this way you can save some effort by eliminating instruments that don't meet your needs in terms of content and that don't pass muster when it comes to reliability and validity.

A checklist of the steps is included, for your convenience, at the end of this volume. There the reader will also find a glossary of many of the technical words used here.

STEPS IN THE EVALUATION PROCESS

The effective evaluation of multi-rater feedback instruments requires the following sixteen steps:

1. Find out what is available
2. Collect a complete set of materials
3. Compare your intended use to instrument characteristics
4. Examine the feedback scales
5. Familiarize yourself with the instrument-development process
6. Learn how items and feedback scales were developed
7. Find out how consistent scores tend to be
8. Assess basic aspects of validity: Does the instrument measure what it claims to measure?
9. Think about face validity
10. Examine the response scale
11. Evaluate the feedback display
12. Understand how breakout of rater responses is handled
13. Learn what strategies are used to facilitate interpretation of scores
14. Look for development and support materials
15. Compare cost: Value for the price
16. Consider length a minor issue

GETTING STARTED

STEP 1: FIND OUT WHAT IS AVAILABLE

The availability of managerial feedback instruments is increasing at a tremendous pace. Today there are as many promising instruments under development as there are good instruments for sale. So your first task should be to gain some knowledge of what is out there in order to choose the best possible sample of instruments to review.

In the short run a good way to familiarize yourself with what is available is to search one of several guides that categorize and

review instruments. *The Directory of Human Resource Development Instrumentation* (Peters, 1985) is one such guide. It provides basic descriptive and technical data on many kinds of instruments available for use in organizations. Other guides include *Mental Measurements Yearbook* (Conoley & Kramer, 1989; Mitchell, 1985); *Tests: A Comprehensive Reference for Assessment in Psychology, Education and Business* (Sweetland & Keyser, 1986, 1990); and *Psychware Sourcebook* (King, 1987, 1988). These can usually be found in the reference section of libraries. Over time, it may be useful as well to keep a file of the instrument brochures you obtain, since many of the directories are not published often enough to keep you updated on the very newest products.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS

When you have identified several instruments that you wish to evaluate, you need to obtain five pieces of information about each of them. You cannot make an informed choice using only a copy of the instrument or a promotional brochure.

Specifically, for each instrument you wish to consider, you should obtain the following:

- A copy of the instrument itself. If the instrument has one form for the subject to rate himself or herself and a separate form for the others who will rate him or her, get both.

- Sample feedback (a representation of what the manager will receive after the instrument is scored). You can't tell what type of feedback your managers will actually receive by looking at the instrument they will fill out. The sample could be a complete report or it could be part of a report such as an example of feedback display in a manual for the instrument (e.g., the technical or trainer's manual). Either type will do.

- A technical manual or other official publication which outlines in detail the developmental and psychometric research done on the instrument.

- Information about any supporting materials that accompany the scored feedback, such as interpretive materials, development guides, goal-planning materials, and the like.

- Information about price, scoring, and whatever certification or training may be required to purchase or use the instrument.

It is not at all unreasonable to request this quantity of information. American Psychological Association guidelines require that this information be available upon request when an instrument is offered for sale.

In addition to seeking the recommended amount of high-quality information, you should, through all the steps that follow, look for evidence of a commitment to continuous improvement on the part of each instrument's developer, especially if an instrument is not brand new. As we will discuss in the section on validity, research should always be in progress, because no instrument can ever be considered valid once and for all. Expect revisions in the scales over time; these are often made when additional validation studies have been completed. Expect revision in the presentation of feedback as well; these are often made as the developer learns from the experience of those who have used an instrument. It is not uncommon for good instruments to have more than one copyright date, since even small revisions to content can cause changes in other areas, such as scale weightings or instrument norms.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS
STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS

It is improbable that one instrument can meet the needs of all managers in an organization. Job demands differ somewhat by organizational level, and even at one management level, the skills that are needed for effectiveness may change over time. It is also highly unlikely that one instrument can meet the needs of all kinds of organizations. The requirements and values of leadership or management vary from company to company. Thus, in searching for

an instrument to provide feedback to managers, a person is typically looking for one that will satisfy the needs of a particular population of managers in an organization with specific leadership or management needs.

Although nearly every instrument has a statement of purpose describing the level it targets, there seems to be little relationship between the level of target manager and the domains of activity or behavior assessed (see the "Who and What is Assessed" section of volume II for further discussion of this point). An instrument targeted toward all levels of management might not be right for middle managers in your organization because the competencies assessed are not in line with company-wide management-development goals. An instrument targeted toward executive levels might be right for your middle managers if the competencies assessed are in line with your management-development goals.

More important than considering advertised audience is discovering the norm group, if any, to which managers will be compared. By *norm group,* we mean the group of managers whose scores are stored in the vendor's test database and are output as a comparison group on every individual feedback report. If the norm group is comprised of senior-level managers, whose skills are likely to be more highly developed, the moderate scores of middle managers will look worse than they would if they were compared to other middle managers.

Therefore, look for instruments that have been normed on a sample similar to your target managers; consider level, organization type, and demographics (e.g., ethnicity, gender). And if you do choose to use an instrument whose norm group differs from your target audience, it may be wise to choose one that allows raters to register "Not applicable" as a response to behaviors and skills, so that raters are not forced to evaluate managers on dimensions they see as inappropriate.

But be forewarned: The feedback instruments we are concerned with here have been developed for use in management-development efforts, either in the classroom or in individual feedback settings. There are virtually no instruments in this arena that have been intentionally developed or adequately tested for other purposes—such as making selection or promotion decisions.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS
STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS

STEP 4: EXAMINE THE FEEDBACK SCALES

In evaluating individual instruments you should begin by examining the scales on which feedback will be received. There is a detailed discussion of scales in step 6, but what you need to know at this point is that the set of scales represents the content or competencies on which managers will be measured, and each individual scale represents a slice of managerial work; as a whole the scales provide a portrait of leadership or managerial effectiveness. Using the sample feedback you have obtained, you should consider the following when looking at scales:

- Is your organization wedded to a particular way of representing what it takes to be effective in your business, or do you have a particular model underlying management-development efforts?

- If so, does the range of scales fit with what you see as relevant competencies for managers in your target group?

- Does the number of scales seem reasonable?

- If, in your judgment, an instrument does not have enough feedback scales that seem relevant to your target group, or if it has too many that seem irrelevant, drop it from further consideration.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS
STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4: EXAMINE THE FEEDBACK SCALES
STEP 5: FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS

In order to know how to identify quality when you see it, you must understand the basics of sound instrument development.

The development process can be seen as occurring in four stages:

- developing instrument items and feedback scales;
- assessing reliability and validity;
- designing the feedback display; and
- creating supporting materials.

At each stage different issues are being addressed.

When items and scales (fully defined in step 6 below) are being developed, an author must identify, as much as possible, the full range of behaviors or skills that he or she believes measures management or leadership competency. Another question at this stage is whether items of behavior or competency cluster in groups which are internally homogeneous, distinct from each other, and useful for feedback.

To assess reliability, the author of an instrument must consider whether the measurement of these skills or competency areas is stable in a variety of ways. To assess validity, the author must determine whether the scales really measure the constructs they were intended to measure and whether they predict effectiveness as a manager or leader. Because multi-rater feedback instruments are primarily intended for individual development, the question of whether the areas assessed are developable must also be asked.

When designing feedback, the author aims to maximize the manager's understanding of the data and to enhance the impact of the data on the recipient.

In creating supporting materials, the aim of the author is to help the facilitator and the feedback recipient gain deeper understanding of the theory or research behind the instrument and thereby enhance their ability to interpret and work with the data.

Your task as an evaluator is to assess the work completed in each of these four stages and balance what you find against the needs of your target group.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS
STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4: EXAMINE THE FEEDBACK SCALES
STEP 5: FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6: LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED

Instruments that assess managerial practice or leadership ability are dealing with complicated phenomena. These phenomena cannot be adequately represented by a single behavior or characteristic because they are comprised of many closely related behaviors or skills. To adequately measure these complex competencies, instruments must use scales which are made up of multiple items.

The process of instrument development typically begins with the construction of items which represent behaviors or characteristics believed to be related to effective management or leadership.

Items can come from a variety of places. Sometimes, the author refers to theory (leadership theory, theory of managerial work) to develop specific behavioral statements or statements describing characteristics or skills. At other times items are developed as a result of research, with researchers creating behavioral statements or descriptions of characteristics or skills based on data they have collected from managers. Another way that items can be written is by basing them on the organizational experience of the author(s). People who frequently work in a training or consulting capacity with managers may feel they can capture in a set of statements the essence of the leadership or management effectiveness they have observed.

The better instruments tend to be those that have used a combination of approaches in their development. A basis in theory provides an instrument with a rich set of validation strategies, while research can provide data from working managers. Ultimately, the quality of the final product depends on a combination of

the quality of the theory, research, or experience of the developer; his or her skill in translating theory, research, or experience into written items; and the attention paid to instrument development and feedback design. A complete evaluation on your part will reveal the level of quality at all these stages.

The nature of items can vary, regardless of their origin. Items can be phrased behaviorally (e.g., "Walks around to see how our work is going"), phrased as skills or competencies (e.g., "Is good at influencing the right people"), or phrased as traits or characteristics (e.g., "Highly motivated").

Instrument feedback is usually presented to the target manager as scores on scales (groups of items). Because scales tend to be abstract (e.g., "Charisma" or "Resourcefulness"), it may be difficult for target managers to set goals for change based on this type of data. Thus, to help managers process their data, some instruments provide scores on the individual, more concrete, items which comprise these scales.

If item-level feedback is given, item content is important to consider. Feedback on behavioral items may be easiest for managers to use in setting goals for change because they are the most concrete. Behavioral changes are the easiest for co-workers to see as well. Change on this type item, however, can be the most superficial in terms of enhancing personal development. At the other extreme, feedback on traits or characteristics can be the most difficult to use, and change on this type item can be the hardest to observe. But change on this type item may be more likely to enhance personal development. Feedback on specific skills probably falls somewhere in between these two extremes: It is moderately easy to use when changes are observable; and it involves some real skill development.

The items discussed above are good examples. If one receives a low score on a behavioral item such as, "Walks around to see how our work is going," it will be relatively easy to change (i.e., "Walk around more") but will probably lead to little in the way of personal development for that manager. If one receives a low score on a skill-based item such as "Is good at influencing the right people," it will be harder to change, since the manager will have to find out how to become better and then will need to improve. But the result can be important skill development. Finally, receiving a low score on an item such as "Highly motivated" can be the hardest of all to change. Change will require the manager to reflect and discover

why motivation is low, decide what it will take to feel more motivated, and then to make whatever personal or life changes are necessary. This kind of change, however, can be the more developmental.

If individuals are left on their own to process feedback (i.e., no trainer or facilitator will be available), or if an instrument is not accompanied by comprehensive interpretive and development materials, the impact of item content is critical. The harder items are to interpret, the more difficulty managers will have in coping with the feedback and the more important quantity and quality of support becomes.

Once items are created, instrument development proceeds to the task of constructing the scales on which feedback will be given. Multi-rater instruments assess complex phenomena that cannot be adequately represented by a single behavior or characteristic. Multiple items are grouped together to represent the set of closely related skills or behaviors that make up the leadership competencies (for instance, "Resourcefulness" or "Planning and Organizing"). Actual responses to the items in a scale should cling together to form a coherent package, internally homogeneous and distinct from other scales.

How the scale-development process is conducted is critical, since the resulting scales will form the basis of the model of leadership, management, or effective performance that you will be presenting to your managers. Your task as the evaluator is to discover whether the process of scale construction seems reasonable and complete. To determine that, you will need to look in the technical manual or in published technical papers.

There are typically two aspects of scale development: the statistical and the rational/intuitive. The statistical aspect involves using statistical-clustering procedures, such as factor analysis, cluster analysis, or item-scale correlations to group items into scales based on the degree of similarity in the response patterns of raters (i.e., people rated high on one item are also rated high on other items in that scale). The rational/intuitive aspect involves grouping items together based on the author's expectations or experience about how different skills or behaviors relate to one another. Some instruments have used one of the two processes in developing scales from items and some have used both.

Look for some evidence of statistical analysis such as factor analysis, cluster analysis, or item-scale correlations. Although it is

not important that you understand the workings of the statistical procedures, it is critical to realize that their goal is to reduce a large number of items to a smaller number of scales by grouping items together based on how these behaviors or characteristics co-occur among real managers.

Also look at whether the items grouped together by the statistical techniques make sense. Because statistical techniques group items according to their empirical correlation, items which do not relate conceptually to other items can end up grouped together. For example, "Being attentive to the personal needs of subordinates" may be statistically related to "Maintaining an orderly work space," not because these two skills are conceptually linked but because the managers used in the analysis happened to score high on both. An intuitive look at scale composition can weed out item groupings that make little sense.

On the other hand, if feedback scales appear to have been created solely by intuitive means, be aware that you have no data on how well the behaviors, skills, or traits work together to measure a more abstract construct—the question of whether these item groupings really represent competencies remains unanswered.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS
STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4: EXAMINE THE FEEDBACK SCALES
STEP 5: FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6: LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7: FIND OUT HOW CONSISTENT SCORES TEND TO BE

Information for this step and the next will again be found in the technical material on the instrument. First, look for a section on "Reliability," or several sections on "Test-retest Reliability," "Internal Consistency," and "Interrater Agreement."

Basically, reliability is consistency. Without evidence of reliability, we do not know whether the items and scales of an instrument are solid enough to hold up even under stable conditions

or within homogeneous groups of raters. In other words, you do not know whether the meaning of items is clear enough so that a single rater will rate a manager the same way over relatively brief periods of time; you do not know whether items which are grouped together in a scale are measuring the same competency; and you do not know whether groups of raters who bear the same relationship to a manager tend to interpret the items similarly.

There are three aspects to reliability:

- homogeneity within scales;
- agreement within rater groups; and
- stability over time.

Homogeneity Within Scales: Basic Issues in Internal Consistency

Homogeneity within scales is called *internal consistency*. This type of reliability applies to the scales on which feedback is to be given, rather than to the individual items to which raters respond.

Internal consistency measures are based on the average correlation among items and the number of items in a scale. Fundamentally, it asks whether all the items which make up a single scale are, in fact, measuring the same thing (as inclusion in a single scale would suggest). If they are, then managers who exhibit one of the behaviors that define the scale should also tend to exhibit the behaviors described by other items on that scale. If this coefficient is low, either the scale contains too few items or the items have little in common. If the coefficient for a scale is very high, there may be more items than are really necessary to adequately measure the construct.

Though several statistical procedures exist for testing internal consistency, Cronbach's alpha is the most widely used. An alpha of .7 is generally considered acceptably high.

Agreement Within Rater Groups: Basic Issues in Interrater Reliability

Agreement within rater groups is commonly called *interrater agreement*. This concept assumes that two or more subordinates, peers, or supervisors, having had adequate opportunity to observe a manager's behavior, should tend to agree on performance measured by clearly stated and reliable items or scales.

Interrater reliability does not tend to be very high for feedback instruments for several reasons. First, raters are typically untrained. Trained raters will produce high agreement by using a common evaluative framework. The people usually completing these instruments, however, are each relying on nothing more than their own perceptions, and perceptions tend to vary. Second, a group of subordinates may be managed differently by their manager. Thus, their within-group agreement may be relatively low because the target manager is actually interacting with each of them differently. Third, some behaviors may elicit higher interrater agreement than others. And fourth, interrater agreement among peers may be lowered to the extent that managers interface differently with different kinds of organizational peers.

Again, this type of reliability coefficient does not tend to be very high for feedback instruments; it should not, however, be extremely low. Look for correlation coefficients, within specific rater groups (i.e., peers, subordinates), that are at least .4.

Additional Considerations

It is not uncommon to find instruments that have not been studied for their interrater reliability. There is some disagreement about whether it is necessary or useful to study interrater agreement in multiple-perspective instruments, since the perspectives taken by various raters (peers, boss, subordinates) will differ according to the organizational relationship they have with the target manager. The argument is that these differences of organizational perspective will be reflected in what are then artificially low interrater correlations.

Yet careful analysis can be, and has been, done on agreement within the different perspectives; and when approached this way,

there should be more agreement within rater groups than between them. For example, a single group, say peers, should show higher intergroup agreement than do peers as compared to subordinates. This approach recognizes both the desirability of reasonable agreement among raters of the same type and the reality of disagreement between raters of different types.

In looking at agreement within rater groups, one needs to be wary of a very high correlation, since this may be an indication of raters' overall bias toward the individual (the so-called halo effect), instead of a clear reflection of the individual's performance. Therefore, within-group agreement should not be lower than .4 or much higher than .7.

Stability Over Time: Basic Issues in Test-retest Reliability

Stability over short periods of time is called *test-retest reliability*. It is studied in relation to both items and scales. An item should be stable enough to produce a similar response this week and next week, given that no significant event, such as training, has intervened. If the meaning of an item is unclear or debatable, it may produce varying responses because of this ambiguity, rather than because any real change in behavior or skill level has taken place. Low test-retest results on an item indicate the item should be revised and retested or dropped from a scale.

Adequate test-retest reliability is especially important if you expect to use an instrument before and after training or over time to measure program impact or individual change. If test-retest reliability has not been demonstrated, you cannot know whether changes in scores over time represent real changes in performance or are merely the result of instability in the instrument itself.

Look in the technical materials for test-retest coefficients of at least .4. These coefficients can range from zero to 1.0, with .4 being minimally adequate and .6 to .7 quite high. This coefficient tends to be low for multi-rater instruments because these instruments capture perceptions of behavior or competence and how one is perceived is probably more unstable than either actual behavior or competence level. Yet it is still important that the developer has weeded out of the instrument those items with very low stability over time.

Additional Considerations

A few instrument developers argue that test-retest reliability is an inappropriate area of study, either because they believe simply completing a questionnaire will affect responses on a retest or because their instruments were designed to create change and, therefore, can't be assessed on this criteria. Although both of these assertions may be partly true, neither negates the need to examine test-retest correlations to look at basic item stability.

Responding to a set of test items may, in fact, alert respondents to dimensions of managerial effectiveness of which they had been previously unaware. As a result, raters may rate a manager somewhat differently, even a day later, than they did the first time around, having now had time to observe the desired behavior. That may be another reason why test-retest correlations do not typically turn out higher than .6 or so for multi-rater instruments.

In order to make a meaningful statistical comparison, the appropriate way to conduct test-retest reliability studies is to compare data from two administrations of the instrument which were completed over three to six weeks, with no intervening feedback of results. This method ensures that study participants are unaware of their initial ratings when they complete the instrument for the second time.

Although completing the form the first time may slightly influence the way raters observe managers or influence the way managers rate themselves, managers are not likely to be significantly changed merely by completing the form. Really significant behavioral change is quite difficult to achieve, usually requiring both powerful feedback and intense follow-up.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS
STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4: EXAMINE THE FEEDBACK SCALES
STEP 5: FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6: LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7: FIND OUT HOW CONSISTENT SCORES TEND TO BE
STEP 8: ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE?

The next step is to look for evidence of validity. In the instrument-development process, this is the process of reality testing. It is here that the developer begins to see whether the model holds up for real managers in real jobs.

Studying validity can be seen as involving three components. The first is determining whether the instrument actually measures what it intends to measure. For instance, does it measure leadership rather than intelligence or power? The second is seeing whether what is measured really makes a difference. Are higher scores actually related to greater effectiveness in real organizations? The third component is evaluating whether what is measured can change as a result of training, feedback, or other developmental experience. Can the instrument's scales be shown to represent areas that are amenable to development?

Although the phrasing of promotional materials for some products would lead you to believe that their successful completion of the first component automatically provides evidence for the second or third, these components are actually independent, and they can be progressively more difficult to achieve. But all are critical to accurate interpretation and maximum use of the feedback.

Validity is integrity. If you do not know for sure whether an instrument measures what it says it measures and you do not have evidence that scores are related to effectiveness on the job, you cannot know how to interpret data from that instrument. If you have not seen evidence that managers can improve their perfor-

mance, you cannot know whether the instrument domains represent areas in which managers can improve.

These considerations are especially important because managers typically will be using their feedback to set performance goals or to try to change their behavior, and most will take what others say very seriously. Established validity is also helpful in dealing with managers who receive very negative feedback from others. These managers may, in defense of their self-esteem, attack the integrity or validity of the instrument by arguing that it assesses areas that don't make a difference or are areas over which the manager has little control.

Look in the technical materials for a section on validity. There may actually be several sections labeled "Construct Validity," "Concurrent Validity," and the like. Though these may be useful categories for the statistician, one does not have to be thoroughly versed in the meaning of these concepts to get enough information about the extent of validation research on the instrument. (The interested reader can find a discussion of the meaning of the different categories of validity in the "Additional Considerations" section below.)

The modern view of validity, in fact, seems to be that, rather than examining these several, often overlapping categories of validity, one can focus on the main inferences that underlie the use of an instrument. And by *use* we mean both the use for which it was intended and the use you plan to make of it. The typical assumptions underlying the use of instruments for management development are, as indicated above, that the instrument is measuring what it says it measures, that scores are related to effectiveness, and that the constructs measured are amenable to change.

Look in the validity section of the technical materials for studies which compare scores in the instrument to scores on another validated instrument (i.e., an instrument whose psychometric properties are known). Favorable results in this type of study will provide some evidence that the instrument measures what it intends to measure. Another kind of study to look for would be one that tests hypotheses about observable behaviors. Other types of studies which can test this assumption are described below in the "Construct Validity" section.

The next underlying assumption to be tested is that scores on the instrument are related to important aspects of managerial

effectiveness on the job. The typical way to test this hypothesis is to compare scores on the instrument to some other data relating to effectiveness or competence. These data, if collected at the same time as the ratings on the instrument, reflect current effectiveness. Raters may be asked to rate the target manager on a different effectiveness measure; current performance appraisal data may be collected; or data may be obtained which relate to current employee satisfaction or work-group productivity. In addition, effectiveness data or data on advancement may be collected after some time has elapsed, so that we can know whether performance at a later time can be predicted by scores on the instrument from an earlier time.

Look within any categories you find for studies which relate scores on the instrument to another measure of effectiveness on the job. This other measure could be another instrument which has been completed on the manager, or it could be a rating on effectiveness done by the manager's boss or co-workers. In a correlational study, look for statistically significant correlations which are at least moderate in magnitude (at least .4). Or look for a study which has compared instrument scores for groups of managers considered to be performing effectively or ineffectively—high performers or low performers. In this case, look for a result which shows that scores on the instrument were able to predict membership in the high and low groups (that is, instrument scores of managers in the high group were greater than scores of managers in the low groups).

The third type of study to look for is one that shows that the concepts measured are amenable to change as a result of training, feedback, or other experience. This is the most difficult study to find, since very few instrument developers have done this research. Studies of change resulting from feedback are scarce.

One additional issue in examining validity is the sample used in any of the research studies. How well do the characteristics of the sample reflect those of your target population? If the sample was primarily white male and your target audience will be significantly different, or more diverse, check to see whether gender or ethnic differences have been found in other research.

If you plan to use an instrument internationally, the issues to be concerned with involve the quality of translation, the availability of international norms, and evidence of validity in international populations.

Translating an instrument should be a carefully executed, multistage process, including both translation from English to the

other language, as well as independent translation from the other language back into English. The use of back-translation ensures that meaning has not been lost in the process. Careful attention should be paid to idioms during this process. For example, "Putting out fires" is common jargon among American managers but has a different, more literal interpretation in both the nonmanagement arena and in areas outside the United States.

It is not always the case that an instrument translated for international use includes translated feedback and development materials. Many instrument vendors have translated only the instrument itself but not the feedback report or development guide. Products such as this can be used for English-speaking managers whose raters are not fluent in English. They cannot, however, be used for managers who themselves are not fluent in English because these managers will be unable to read their feedback. Thus, if your target managers do not speak English, check thoroughly the completeness of translation done for all materials related to the instrument.

Finally, instruments developed from research, theory, or experience with American managers must undergo further testing with the population for whom the instrument will be used. Norms must be collected so that cross-cultural comparisons can be made. If managers in the international group score differently than American managers, international norms should be made available for use in other cultures. Also, cross-cultural validity studies need to be undertaken to verify that scores on the instrument are related to effectiveness in the other cultures.

The rate at which instrument publishers have begun to translate instruments for international users has increased, and it is not uncommon to find instruments that have been translated into more than one language. It is, however, difficult to find American-developed instruments that have been validated for the population for which they're translated.

So a cautious approach is warranted when selecting American-developed instruments for non-English-speaking managers. Look for one that has been carefully translated and which has been studied with respect to cross-cultural differences in norms, factor structure or scale reliabilities across different groups, and whether scores are related to effectiveness in similar ways across different cultures.

Keep in mind, however, that no instrument is ever proven valid once and for all. Validity is a matter of collecting evidence over time in varying situations for different populations. In fact, any time the intended use of an instrument departs significantly from the assumptions underlying validity research to date, more research should be done.

It is useful to think of validity as a continuum ranging from no evidence that the inferences underlying the instrument are valid, to some evidence, to a good mix of evidence among varying groups and in different contexts. In general, the number of validity studies done should correspond roughly to how long the instrument has been around. Newcomers will have fewer studies to report, but all instruments which offer feedback from co-workers should have at least one study completed which indicates the instrument is measuring what it claims to measure and one which shows that scores are related to effectiveness.

Additional Considerations: Ways of Defining Validity

Validity has been carved up by psychometricians and statisticians in a variety of ways (content, construct, concurrent, predictive).

Content Validity. Content validity has to do with the degree to which the items are a representative and comprehensive measure of the phenomenon in question, in this case leadership or management skills. For example, a final exam for a college course should be made up of test questions that do not go beyond the material covered but are comprehensive, in that they do not represent too narrow a slice of the material.

The issue of representativeness (the question of whether items relate closely to the larger concept or concepts being measured) can be studied by examining the correlations between the individual items which comprise each scale and the overall scale scores. Although high item-scale correlations indicate that the chosen items do indeed relate closely to the larger concepts being measured, this type analysis does not tell us whether all measurable aspects of the larger concept (e.g., management or leadership) are being covered by existing items. This aspect of content validity is often more a matter of professional judgment than statistical analysis.

Construct Validity. A construct is an abstract variable that cannot be directly measured. Temperature and humidity are not constructs; but leadership, resourcefulness, and integrity are.

Construct validity refers to the ability of the instrument to measure phenomena that are hypothesized to exist but for which we can have no real-world measure. Every study of construct validity represents a test of hypotheses about relationships between a set of real behaviors. Some constructs are small or narrow, in the sense that they may be completely described by a few behaviors. Other constructs are so large or broad that they may never be perfectly defined by even a very large number of behaviors. Broad constructs, like leadership, are usually measured using only a subset of all possible behaviors.

Thinking about construct validity begins with thinking about the theory, if any, that has been used to specify the meaning of the construct; how the construct should be related to other constructs; and how it should be related to specific observable behaviors. One way construct validation research can proceed is by testing hypotheses generated about observable behaviors. Another strategy may be to compare scores on the target instrument to scores on a similar instrument whose psychometric properties are known.

If, as with management and leadership instruments, several constructs (such as resourcefulness or integrity) are hypothesized to make up the overall leadership or management construct, then another research strategy could be employed. This strategy would enable one to investigate whether items within one scale are more closely related to each other than they are individually to items comprising other scales.

Since constructs cannot be directly measured, another way to approach construct validity is through a multimethod study. For example, if we are measuring more than one construct (resourcefulness and integrity) and can measure them in multiple ways (paper-and-pencil and rater observations), then we can discern how much of the correlation between those two constructs is due to the measurement method and how much is due to overlap in the constructs themselves. To obtain evidence of construct validity in our example, the correlation between the paper-and-pencil and rater observation measures of resourcefulness should be greater than the correlation between resourcefulness and integrity using either of the two measurement techniques. *Convergent validity* is the label given to the

condition of high correlation between two different measures of the same construct.

Concurrent and Predictive Validity. The next two types of validity, concurrent and predictive, both have to do with using scores to predict some other attribute(s). With managerial feedback instruments, we are concerned in both instances with making predictions about performance or effectiveness.

Both types are temporal. Concurrent validity has to do with the relationship between scores on an instrument and performance or effectiveness measured at approximately the same time the instrument was completed. The relevant question is: How accurate is the instrument's assessment of current performance or effectiveness? Predictive validity has to do with the relationship between scores on the instrument measured at an earlier time and performance or effectiveness sometime in the future. (Though some people use the term *predictive* to cover any study which predicts effectiveness, we will limit our use of this term to mean only those studies which relate current scores to future performance.) The relevant question here is: How accurately can the instrument predict future performance or effectiveness?

Both concurrent and predictive studies are relevant for instruments used for management development, because in helping managers become more effective we are interested in both current effectiveness and effectiveness over the long term.

Criterion issue. In studies of both these types, the nature and measurement of the criterion variable (i.e., the measure to which scores on the instrument will be compared) is a key issue. Examining the criterion measure employed to validate an instrument can be as important as examining the study's findings.

Remember that instruments are valid for very different reasons. All leadership and management instruments are intended to assess factors associated with success or effectiveness. But success can be thought of as forward movement in one's career (i.e., promotion upward through the organization) or it can be thought of in terms of effectiveness as a leader or manager (i.e., how satisfied others are with the individual's performance, how well the manager's subordinates perform, measures of profit-and-loss, and the like). These two phenomena are not necessarily the same (Luthans, 1988). Understanding how success has been defined in the research on an instrument and knowing how that compares

with your own intended use of it is key to interpreting the results of any validity study.

The main, and only really appropriate, use of data from any multi-rater feedback instrument is for management development or assessment for development. No instrument of this type has been intentionally developed or significantly tested for other uses (e.g., selection or promotion). Managerial feedback instruments that compare self-view to the views of others are basically dealing with perceptions of behavior or competence, rather than competence measured in any verifiable sense. An individual or organization planning to use an instrument of this sort for selection or promotion purposes is in treacherous territory and will need to become aware of professional standards and government guidelines not covered in this report.

For management development, the main inferences to be drawn are that high scores are related to effective current performance and that the competencies represent areas managers are able to develop.

Finding an appropriate criterion variable for a study of the first inference would involve choosing a definition of effective performance: If it is defined as the ability to effectively mold a high-performing team, the criterion variable might be a measure of team performance; if it is defined by the boss's perception of the individual's performance, the criterion variable might be promotability ratings or performance-evaluation data; and if it is defined by subordinates' job satisfaction, the criterion variable might be rated "effectiveness," a job satisfaction scale, or a measure of how well the manager is meeting the work-related needs of subordinates. Variables less directly related to managerial effectiveness, such as bottom-line business profitability, may not be the best criterion measures. Businesses can be profitable for reasons other than good management, and businesses can fail despite good management.

A useful study of the second inference would show that scores improve as a result of training, feedback, or other development and effort on the part of the manager receiving the feedback. As mentioned earlier, this type study is the more difficult to find.

Choosing the criterion variable for a validity study is one of the more problematic but critical phases of the research. The choice should be made with reference to the main inference(s) to be drawn in using the instrument. In general, criterion data from an independent source (i.e., a source different from the person or persons

providing ratings on the instrument) is usually considered better data because the independent source reduces the natural correlation (or bias) occurring when both pieces of data come from the same person.

Reliability of criterion measure. A good criterion measure should also have a high degree of reliability. In other words, the success that it measures should be constant over time. When the criterion measure used is a rating of effectiveness, it should be subject to some of the same concerns about reliability as the instrument itself. That is, raters' judgments about overall effectiveness or their own expressions of satisfaction with their leadership may be stable or unstable over short periods of time. Although this issue is an important one to think about in evaluating the quality of a validity study, it will be difficult to find any such research because very few instrument developers have taken the time to study the reliability of their criterion measures (see the "Validity" section of volume II).

There is No Perfect Instrument: Balance is the Key

Remember throughout this process that there is no perfect instrument—no one instrument that is perfect for everyone, no one instrument that will always be ideal for you, and no instrument that couldn't stand some improvement. As solid and powerful as some are, each feedback tool has its weaknesses as well as its strengths.

The key to good instrument selection is balance—knowing what you need and seeking a balance between your needs and quality standards. You may have already made some conscious trade-offs and before this selection process ends, you may make some more—sacrificing one good to get another. But to maximize the value of your choice—to most effectively use the instrument you choose—you need to be aware of the trade-offs you have made—what you have lost and what you have gained.

Keep in mind that no instrument is proven valid once and for all. An instrument having several good studies defining effectiveness in varying ways is stronger than one that has only a single study or several studies of effectiveness defined the same way.

CONSIDER POTENTIAL IMPACT

Now that you have reviewed the origin of items and feedback scales and have looked over the instrument in terms of reliability and validity, you have available an instrument or set of instruments that you know have been constructed according to sound psychometric principles. At this point you can begin to evaluate and compare instruments in terms of their potential impact—what kind of feedback will be provided and how that will be delivered.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS
STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4: EXAMINE THE FEEDBACK SCALES
STEP 5: FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6: LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7: FIND OUT HOW CONSISTENT SCORES TEND TO BE
STEP 8: ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE?
STEP 9: THINK ABOUT FACE VALIDITY

Go to the instrument itself—the list of items to which the manager will respond. Read through the items carefully. Experiment with them by trying to apply them to a manager whose skills you can observe. Do the items make sense when applied to someone you know? If you go through this exercise with several different managers in mind, you will get a sense of how managers who complete the instrument will feel about its face validity. If the items are awkward or hard to apply, think about dropping this instrument from the group under your consideration.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS
STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4: EXAMINE THE FEEDBACK SCALES
STEP 5: FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6: LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7: FIND OUT HOW CONSISTENT SCORES TEND TO BE
STEP 8: ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE?
STEP 9: THINK ABOUT FACE VALIDITY
STEP 10: EXAMINE THE RESPONSE SCALE

Look at the main response scale—the choices people will have when they respond to each item. Examining this helps you understand how people will be asked to think about the competencies represented by the scales. Will they be asked to rate the competencies as either "Strengths" or "Areas for development" or will they be asked to make finer distinctions in rating levels of skill?

Response scales may range from 2 points to 7 (or more) points. Both shorter and longer scales have their advantages as well as their disadvantages. A 2-point scale will force one to choose between alternatives such as "Yes/No," "Strength/Development needed," or "Do more/Do less." The advantage of the shorter scale is simplicity of feedback.

The trade-off with a shorter scale is in data richness. With a short response scale, raters will not be able to report mid-range levels of competency and may fail to respond to items if they are unwilling to say definitely "Yes" or "No," definitely "Strength" or "Development needed." And short scales are not useful if one intends to use a pre- and post-test design for looking at the impact of feedback or training, because these will fail to pick up the subtle changes fostered by even the best training or change efforts. If any follow-up studies will be conducted down the road, an instrument with a longer scale will be needed.

A longer scale (5 to 7 points) will ask raters to make finer distinctions. The trade-off with a longer scale may be some ambigu-

ity in the meaning of what may be too many response categories. Choosing between "Occasionally" and "Sometimes" may be difficult on some items or for some managers. Some examples of longer scales include:

Never	Very uncharacteristic	Among the best
Occasionally	Uncharacteristic	Better than most
Sometimes	Slightly uncharacteristic	Average
Usually	Slightly characteristic	Worse than most
Always	Characteristic	Among the worst
	Very characteristic	

As can be seen in the three examples above, some response scales give raters a middle-of-the-road choice, such as "Average," whereas others force raters to take at least a slightly positive or slightly negative position.

In addition, a response scale may have "Don't know" or "Not applicable" alternative choices. These may be useful if you are using the instrument for a combined population of managers and non-managers—or with people new to their positions—and want people to have these rating options. The downside may be that raters overuse the "Don't know" option to avoid making difficult rating choices.

STEP 1: FIND OUT WHAT IS AVAILABLE
STEP 2: COLLECT A COMPLETE SET OF MATERIALS
STEP 3: COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4: EXAMINE THE FEEDBACK SCALES
STEP 5: FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6: LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7: FIND OUT HOW CONSISTENT SCORES TEND TO BE
STEP 8: ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE?
STEP 9: THINK ABOUT FACE VALIDITY
STEP 10: EXAMINE THE RESPONSE SCALE
STEP 11: EVALUATE THE FEEDBACK DISPLAY

Type of Display

There are two types of feedback display: graphic display and narrative display. Many feedback displays use a combination of styles to present data.

Graphic displays are charts or graphs which provide a visual picture of how managers view themselves in comparison to their raters and the instrument norms.

Narrative display is computer-generated feedback in text format rather than graphic format. It is both a presentation and a personalized interpretation of the scored data, as opposed to boilerplate recommendations for change. The text produced often incorporates some of the same strategies for framing feedback as can be found in graphic displays (see the section below on strategies for facilitating interpretation and change).

Narrative interpretation differs from development materials (discussed on page 36) in that narrative is personalized, based on the manager's individual feedback and output as a result of the scoring process. Development materials are off-the-shelf manuals, workbooks, and readings.

Clarity of Feedback Display

Open and review the sample feedback as if it were your own. It is probably best not to use your own feedback (if you've already taken the instrument) in this step, since your feelings about your own results may impede an objective analysis.

Carefully read through the instructions and look at the display. Can you understand it? How easy is this for you? Your own experience in trying to interpret the feedback is a good measure of how others will experience it. If it takes you several attempts to understand the graphic display, it is probably too complicated.

Although people should ideally never be handed instrument feedback without some guidance, it is not always possible to provide large groups of managers with intensive support. In general, the more interpretation and goal-setting that managers will have to do for themselves, the more you should be concerned about the clarity of the feedback display.

STEP 1:	FIND OUT WHAT IS AVAILABLE
STEP 2:	COLLECT A COMPLETE SET OF MATERIALS
STEP 3:	COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4:	EXAMINE THE FEEDBACK SCALES
STEP 5:	FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6:	LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7:	FIND OUT HOW CONSISTENT SCORES TEND TO BE
STEP 8:	ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE?
STEP 9:	THINK ABOUT FACE VALIDITY
STEP 10:	EXAMINE THE RESPONSE SCALE
STEP 11:	EVALUATE THE FEEDBACK DISPLAY
STEP 12:	**UNDERSTAND HOW BREAKOUT OF RATER RESPONSES IS HANDLED**

By *rater responses,* we mean the responses that come from the managers' subordinates, peers, and bosses. Find out how rater data are treated in the feedback display.

Some instruments combine the ratings of all "others" responses into one averaged score. Other instruments will display the averaged rating from subordinates separately but will average together boss feedback with peers. Still others present 360-degree feedback—presenting boss, peer, and subordinate ratings as three separate averaged scores.

Using 360-degree feedback may be more powerful, in that it gives the manager more information. It can also be more complicated to interpret, more compromising of rater anonymity, and more painful to receive. Again, consider how much guidance and support you, or others in your organization, will be providing managers when they receive their feedback. The more guidance people receive, the better they are able to benefit from the enhanced impact that 360-degree feedback can provide.

Another issue with 360-degree feedback is that rater confidentiality may be compromised. This is certainly true in the case of separate feedback from the boss. If only one superior completes the instrument on a manager, his or her responses will be undisguised under the "boss" category. Even if superiors have been made aware (as they should be) that their responses will be known to the manager, confidentiality is still an issue to consider.

Raters who know or believe that their responses will not be confidential may not be completely open in their evaluations, compromising the quality of their data. Although breakout of rater data increases impact, it may decrease confidentiality and, occasionally, data quality.

STEP 1:	FIND OUT WHAT IS AVAILABLE
STEP 2:	COLLECT A COMPLETE SET OF MATERIALS
STEP 3:	COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4:	EXAMINE THE FEEDBACK SCALES
STEP 5:	FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6:	LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7:	FIND OUT HOW CONSISTENT SCORES TEND TO BE
STEP 8:	ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE?
STEP 9:	THINK ABOUT FACE VALIDITY
STEP 10:	EXAMINE THE RESPONSE SCALE
STEP 11:	EVALUATE THE FEEDBACK DISPLAY
STEP 12:	UNDERSTAND HOW BREAKOUT OF RATER RESPONSES IS HANDLED
STEP 13:	**LEARN WHAT STRATEGIES ARE USED TO FACILITATE INTERPRETATION OF SCORES**

Enhancing performance or changing behavior is difficult, even when feedback on performance has been comprehensive. The difficulty is partly because feedback can be overwhelming, especially if there is a lot of it, if it is negative, or if it comes from many different perspectives. Without some road map to or guidance about what is most important to direct one's attention to, the recipient can get lost, ignore important data, or give up on processing the information.

Rather than just presenting self and rater scores, an instrument should build some kind of framework for corrective action into the feedback display.

There are at least seven strategies for helping managers understand their feedback:

- comparison to norms;
- highlighting largest self/rater discrepancies;
- item-level feedback;
- highlighting high/low scores;
- comparison to ideal;

- importance to job or success; and
- "Do more/Do less" feedback.

These can be built into either graphic or narrative displays.

Any feedback display should use at least one of these strategies because without any, the manager has no way of knowing what to make of his or her scores. A better feedback display will use a combination of two or more strategies.

Again, past a point, more may not be better. If well-laid out and clearly integrated, the use of more strategies may increase the meaningfulness as well as the impact of the data. This is because it will give the manager many perspectives from which to view the information and make it more difficult to dispute negative feedback. Yet, for some managers, more information may be too much, especially if the display is comprehensive but not clear. Again, it is up to you to decide whether the balance between information/impact and clarity/simplicity is right for your target group.

Comparison to Norms

Comparison to norms is probably the most widely used of all strategies (see the "Feedback Delivery Strategies" section of volume II). Norms often represent an average for all people who have taken the instrument but sometimes are based on one specific target group (middle managers, for example). These averages are presented for each scale on which feedback is received so that the manager can compare his or her own scores to the norm. The manager uses this information to ask himself or herself: How am I doing compared to others? The more the norm group resembles the manager, in terms of managerial level, the more this comparison will be a relevant one.

Norms can be displayed in a variety of ways. They can be listed as a separate column of data to be compared visually to one's own score, or one's own score can be plotted against norm data. When this latter strategy is used, the graphic display contains standard scores or percentile data. A standard score is a mathematical computation that allows individual scores to be compared to a total distribution of scores by taking into account the mean and standard deviation of the total population. A percentile ranking represents the percentage of persons who have lower scores than the individual being ranked.

Highlighting Largest Self/Rater Differences

Although, by definition, all self/rater feedback instruments present self ratings compared to ratings by others, some feedback displays go beyond the simple presentation of these comparisons to highlight scores on scales where self/rater discrepancies are significant. Managers often wonder how much of a difference between ratings is enough to pay attention to, and this type of highlighting lets them know. It can be very useful, especially when combined with rater breakout of data. Self/peer discrepancies on scales that concern managing people may not be significant, for example, whereas subordinates may rate the manager significantly lower than the manager has rated himself or herself, because supervisory skills are often most visible to those supervised.

Item-level Feedback

Though item-level feedback can be unreliable, it can be useful when used in conjunction with feedback on scales because the items are more concrete. Learning that you score low on "Resourcefulness," for example, is not necessarily helpful unless you know what behaviors or specific skills are linked to "Resourcefulness." The items give you that information. As such, they can give the manager some leads about what to begin to do differently in order to be more effective.

Highlighting High/Low Scores

High and low scores are sometimes presented on individual items and sometimes on scales. This type data can be presented as graphs or as listings of the highest or lowest scores with symbols to denote highest or lowest.

When receiving a lot of feedback, the highlighting of high and low scores can provide a quick overview of participants' strengths and weaknesses. The danger is that when high and low scores are item scores, what is produced in listings such as these can be a smattering of items from many different scales—which is thus a very unreliable summary of the data.

Comparison to Ideal

Comparison to an ideal is not widely used (see the "Feedback Delivery Strategies" section of volume II). When used, it may represent a theoretical ideal for leadership or management, or it may be data provided by the manager himself or herself (e.g., What kind of leader do I see as ideal?) or by raters (e.g., What kind of leader do raters see as ideal?). Self or rater data on actual performance can then be compared to data from self or raters on the ideal leader.

Importance to Job or Success

Importance data may reflect how important a skill or practice is to effectiveness in one's job or to longer-term success in one's organization. The data are usually collected from managers or from their raters when the instrument is initially completed. Sometimes raters rate each item or scale in terms of its importance (to the job or to success), and sometimes they choose a certain number of items and scales that are most important.

The use of this strategy can be very powerful in prioritizing which parts of the feedback may need greatest attention, especially when the manager and his or her boss agree. Managers who rate low on several scales can check the importance data and focus most of their efforts on developmental needs that are important for their short-term or longer-term effectiveness. When the manager and boss do not agree, importance data can provide a fairly non-threatening way to begin a conversation about the overall results.

Some caution should be used, however, in working with data on importance to job or success. Importance data are perceptual data and may reflect people's impressions of what has gotten executives to the top as much as real and critical competencies. Ratings may vary depending on the perspective of the person providing the data (the manager's boss, peers, or subordinates). Mid-level managers may rate importance for success differently than executives. And ratings may change over time with changes in organizational culture or business strategy. In addition, few instruments providing this type feedback have assessed the reliability of these ratings (see the "Feedback Delivery Strategies" section of volume II).

"Do More/Do Less" Feedback

"Do more/Do less" feedback is not used by many instruments but can be a valuable piece of information, especially when the instrument's response scale demonstrates how frequently a behavior or practice is observed. Getting a high rating on such a scale usually means you engage in a particular type of behavior frequently. "Monitoring the progress of subordinates' work" is a good example. A high score would mean the manager monitors work frequently, or does a lot of monitoring. What the manager cannot know from this is whether it is good to do a lot of monitoring or whether he or she is doing the appropriate amount. Although norms will tell the manager whether he or she is doing more monitoring than other managers, norms will not shed light on whether subordinates are satisfied with this level of the behavior. Asking raters to indicate whether managers should do more or do less provides an opportunity to see this side of the picture.

Linking Strategies Together: Issues in Considering Impact of Feedback

Different combinations of strategies will produce different kinds of impact on the manager receiving feedback and may result in different conclusions being drawn from the data. Consider the blend of strategies presented in the feedback design, what that blend will highlight for managers, and what the impact will be. Again, maximum impact may not always be desired. The target should be meaningful feedback with appropriate impact.

If you still have several candidates once you've gotten to this point in the selection process, it is worthwhile to take these instruments yourself, have them scored, and receive your own feedback. A second-best alternative is to fill one out, as an exercise for yourself only, on someone else. This kind of experience is especially important if you have any doubts about the impact the feedback will have or whether the depth of intervention will be appropriate. As you complete the questionnaire itself, consider the feel or nature of the items. Sometimes items which look just fine become unclear or awkward when you actually have to use them to rate a real person.

If you do have the opportunity to review your own feedback, pay attention to your own reactions to the competencies being measured by the scales, how rater data are broken out, and what graphic strategies are used in the feedback display to help managers understand and interpret their scores.

STEP 1:	FIND OUT WHAT IS AVAILABLE
STEP 2:	COLLECT A COMPLETE SET OF MATERIALS
STEP 3:	COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4:	EXAMINE THE FEEDBACK SCALES
STEP 5:	FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6:	LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7:	FIND OUT HOW CONSISTENT SCORES TEND TO BE
STEP 8:	ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE?
STEP 9:	THINK ABOUT FACE VALIDITY
STEP 10:	EXAMINE THE RESPONSE SCALE
STEP 11:	EVALUATE THE FEEDBACK DISPLAY
STEP 12:	UNDERSTAND HOW BREAKOUT OF RATER RESPONSES IS HANDLED
STEP 13:	LEARN WHAT STRATEGIES ARE USED TO FACILITATE INTERPRETATION OF SCORES
STEP 14:	**LOOK FOR DEVELOPMENT AND SUPPORT MATERIALS**

Support for the Manager's Development

At this point in the process, having considered potential impact, you are now ready to consider what support will be available with each instrument for both the manager receiving feedback and the trainer or facilitator responsible for one-on-one consultation or classroom training.

During a feedback process, managers are led carefully through the graphic presentation of their scores and they learn something about their strengths and development needs. Yet they

may remain unaware of what these scores mean. They will have an answer to the question, "What are my strengths and weaknesses as a leader?" but may lack answers to, "So what does that mean?" and "Now what should I do?" With feedback instruments, the "So what?" question can be addressed by interpretive development material and the "Now what?" question by a goal-planning process. These kinds of support materials accompany the feedback packet with some instruments but are not a part of some others.

Instruments can have exquisite accompanying materials ranging from nicely bound development guides for managers (including in-depth interpretive materials, suggestions for development, and goal-planning guides), to notebooks, manuals, overhead materials, extra norm data, and group profile options for trainers. Some instruments, however, offer very little besides the feedback packet.

Although support materials are useful, it is intentional that considering what an instrument offers in this regard is one of the last steps in our evaluation process. It would be a mistake to select an instrument because of its glitzy package without considering the integrity of its development and the quality of its feedback display. Once you know you've gotten the best instrument for you in those two arenas, you may want to choose the one with the best accompanying materials.

And, with supporting materials, quantity is not necessarily the same as quality. More is not necessarily better. For example, for most audiences, it is best to look for materials that help managers come up with solutions and goals of their own, because people are more motivated to implement their own ideas than they are to do what they have been told. Yet some development guides are very prescriptive. They have a "Do A, B and C" approach to development. And some materials appear to be presented in a boilerplate fashion. Sophisticated managers may lose faith in an instrument that looks like it gives everyone the same advice.

Trainer Support

To evaluate support for the trainer, what is available must be compared to your own needs as a trainer or facilitator (or the needs of those who will be responsible for feedback facilitation). Some instruments come with training guides which give a more in-depth

understanding of how to interpret scores on the instrument. Find out if attendance at a certification workshop is required for purchase and use. Supplemental norms may be available, either in the form of organizational norms or norms on unique subsamples (e.g., health-care workers, cross-cultural groups, student leaders, high-level executives). Videos may be available for classroom use, as might other supplemental materials including transparencies, handouts, or supplemental technical research reports.

Scoring Process

Manual scoring (hand-scored) and computer scoring are common methods of producing feedback. Computer scoring can be done by the vendor; licensed scoring programs; or self-administered, self-scoring computer software.

Manual Scoring

Hand-scored instruments allow managers to compile their own data and produce their own feedback. This scoring method gives managers a firsthand knowledge of how scores are calculated and plotted, and it allows for immediate turnaround of feedback to the manager.

Although hand-scoring has these benefits, its strengths can easily become its weaknesses. Asking inexperienced users to score their own instruments increases the likelihood of error and invalid results.

Assuring rater confidentiality is another difficulty with hand-scored instruments. Some instruments require managers to transfer raters' scores onto summary sheets. It can be possible for managers to guess from the number of respondents, their handwriting, or the pattern of responses who each rater might be. Such a compromise of confidentiality can damage the quality of the data for that manager by reducing the degree of honesty in raters' evaluations. In cases where rater confidentiality is of a particular concern, an independent individual can be chosen to hand-score instruments.

Computer Scoring

In general, although all computer scoring tends to increase the total price to the consumer, this method has many benefits. The depth and quality of computer-generated feedback tends to be high. Instead of the manager shifting through mounds of data to find, for example, the ten highest rated items, the computer program can sort and arrange the information in a fraction of the time. Scoring errors are reduced to a minimum by scoring programs which can also take human completion errors into account (i.e., questions answered twice or out of order) and can correct or reject inaccurate data. Using a computer-scoring method to produce feedback also can provide raters with a feeling of greater anonymity and can enhance the quality of their responses because raters know that target managers do not see their rating forms. Computer scoring can be done by the vendor or software can be purchased.

Vendor Scoring. Returning instruments to the source where they were purchased is a very common form of instrument scoring. It does, of course, increase the cost per instrument, as well as introduce the issue of turnaround time (the time it takes to mail forms to the vendor, have those scored, and have feedback reports produced and returned to the client). Though there may be problems, many managers and trainers find vendor scoring to be the most expedient method because it can reduce managers' and trainers' time by relieving them of the job of hand-scoring the instruments. It also tends to produce some of the highest quality feedback in a visual sense.

Some vendors license their scoring programs. Initial costs of the license can be high, but for organizations that purchase large quantities of instruments, this method can be cost effective.

Licensed scoring arrangements will vary according to the clients' needs and work environment. Organizations that have personal computers may be able to purchase software to score instruments on-site or may be able to use both on-site scoring and vendor assistance through remote scoring (using a modem to send raw data to the vendor to score). Computer programs for mainframes can sometimes be purchased through the vendor. If your organization has the equipment (i.e., modem, optical scanner, personal computer, and printer) and the resources to maintain these, this scoring method can provide quick turnaround of a high-quality report for consumers who process a large number of instruments.

Computer Software. Instruments that are sold for use with personal computers usually have scoring programs built into the software. The quality of the feedback can be as high as any other computer-scored instrument. This method can, however, have a benefit over other computer-scoring methods, in that the software can be reused over time. The cost of the software per manager will be higher than vendor- or hand-scored instruments but can be used for as long as the life of the disk.

Each of these scoring methods has its advantages and disadvantages. Time, costs, equipment, frequency of use, quantity and quality of feedback, and confidentiality are all issues that require your consideration before deciding which method best suits your organizational needs.

OTHER CONSIDERATIONS

STEP 1:	FIND OUT WHAT IS AVAILABLE
STEP 2:	COLLECT A COMPLETE SET OF MATERIALS
STEP 3:	COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4:	EXAMINE THE FEEDBACK SCALES
STEP 5:	FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6:	LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7:	FIND OUT HOW CONSISTENT SCORES TEND TO BE
STEP 8:	ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE?
STEP 9:	THINK ABOUT FACE VALIDITY
STEP 10:	EXAMINE THE RESPONSE SCALE
STEP 11:	EVALUATE THE FEEDBACK DISPLAY
STEP 12:	UNDERSTAND HOW BREAKOUT OF RATER RESPONSES IS HANDLED
STEP 13:	LEARN WHAT STRATEGIES ARE USED TO FACILITATE INTERPRETATION OF SCORES
STEP 14:	LOOK FOR DEVELOPMENT AND SUPPORT MATERIALS
STEP 15:	**COMPARE COST: VALUE FOR THE PRICE**

Cost is always a consideration, but what is included in the price can vary so extensively that selecting an instrument based solely on cost can result in your choosing a less than optimal instrument for your use.

To adequately conduct the feedback process for one manager, you would need one self-report instrument, at least three instruments for raters (usually the minimum to have rater data scored), a development guide, and an arrangement for scoring. Sometimes the cost includes all these and sometimes it includes only single copies of the instrument. Certification workshops, organizational profiles, trainers' guides, introductory videos, and post-assessment may all be available but at additional cost. It takes thorough investigation to determine what is included in the quoted price and what supplemental services may be available at additional cost.

STEP 1:	FIND OUT WHAT IS AVAILABLE
STEP 2:	COLLECT A COMPLETE SET OF MATERIALS
STEP 3:	COMPARE YOUR INTENDED USE TO INSTRUMENT CHARACTERISTICS
STEP 4:	EXAMINE THE FEEDBACK SCALES
STEP 5:	FAMILIARIZE YOURSELF WITH THE INSTRUMENT-DEVELOPMENT PROCESS
STEP 6:	LEARN HOW ITEMS AND FEEDBACK SCALES WERE DEVELOPED
STEP 7:	FIND OUT HOW CONSISTENT SCORES TEND TO BE
STEP 8:	ASSESS BASIC ASPECTS OF VALIDITY: DOES THE INSTRUMENT MEASURE WHAT IT CLAIMS TO MEASURE?
STEP 9:	THINK ABOUT FACE VALIDITY
STEP 10:	EXAMINE THE RESPONSE SCALE
STEP 11:	EVALUATE THE FEEDBACK DISPLAY
STEP 12:	UNDERSTAND HOW BREAKOUT OF RATER RESPONSES IS HANDLED
STEP 13:	LEARN WHAT STRATEGIES ARE USED TO FACILITATE INTERPRETATION OF SCORES
STEP 14:	LOOK FOR DEVELOPMENT AND SUPPORT MATERIALS
STEP 15:	COMPARE COST: VALUE FOR THE PRICE
STEP 16:	**CONSIDER LENGTH A MINOR ISSUE**

Time is often a factor when choosing between one instrument with a few and another instrument with a large number of questions. Some managers may find an instrument that requires 45 minutes to an hour to complete a bothersome task and may reject its use. Though a longer instrument may require more initial completion time, the depth of feedback it produces can be worth the initial investment.

CONCLUSION

Congratulations! You have now completed the process and are left with one or more carefully developed and well-designed feedback instruments which best meet the needs of your organization.

You have probably made some compromises along the way. Be aware of what those are and think through their implications. Doing this may help you make the final choice, if you are left at this point with more than one candidate. You may also want to re-contact vendors for any additional information that can help you choose between instruments or aid in the understanding or use of your instrument of choice. And remember to remain aware of the implications of any trade-offs you have made.

REFERENCES

Conoley, J. C., & Kramer, J. J. (Eds.) (1989). *Mental measurements yearbook.* The Buros Institute of Mental Measurements (10th edition). Lincoln, NE: The University of Nebraska.

King, S. (1987-1988). *Psychware sourcebook.* Test Corporation of America: Kansas City.

Luthans, F. (1988). Successful vs. effective real managers. *The Academy of Management Executive, II*(2), 127-132.

Mitchell, J. (Ed.) (1985). *Mental measurements yearbook.* The Buros Institute of Mental Measurements (9th edition). Lincoln, NE: The University of Nebraska.

Morrison, A. M., McCall, M. W., & DeVries, D. L. (1978). *Feedback to managers: A comprehensive review of twenty-four instruments* (Technical Report No. 8). Greensboro, NC: Center for Creative Leadership.

Peters, D. (1985). *Directory of Human Resource Development Instrumentation.* San Diego, CA: University Associates.

Sweetland, R., & Keyser, D. J. (Eds.) (1986). *Tests: A comprehensive reference for assessment in psychology, education and business* (2nd edition). Austin, TX: Pro-ed.

Sweetland, R., & Keyser, D. J. (Eds.) (1990). *Tests: A comprehensive reference for assessment in psychology, education and business* (3rd edition). Austin, TX: Pro-ed.

GLOSSARY OF TERMS

Behavioral item – A word, phrase, or sentence that describes some aspect of behavior.

Cluster analysis – A statistical analysis that groups like objects or items.

Coefficient – A number, resulting from statistical analysis, which indicates the strength of relationships.

Concurrent validity – The relationship between scores on an instrument and performance or effectiveness measured at about the same time the instrument was completed.

Construct – An abstract variable that cannot be directly measured.

Construct validity – The ability of the instrument to measure phenomena that are hypothesized to exist but for which we can have no real-world measure.

Content validity – The degree to which the items are a representative and a comprehensive measure of the phenomenon in question.

Correlation – A numeric indication of the strength of relationship between two variables.

Correlation coefficient – See *Correlation*.

Criterion variable – A behavior or phenomenon separate from, but presumably related to, ratings produced by the instrument.

Cronbach alpha – A measure of the internal consistency of scales, based on the average correlation among items and the number of items; if this coefficient is low, either the scale contains too few items or the items have little in common.

Empirical – Based on experience, observation, or data.

Face validity – Whether the instrument looks like it measures what it is intended to measure.

Factor analysis – A set of statistical approaches to clustering items or variables into groups.

Feedback – Information about a person's performance or behavior, or the impact of performance or behavior, which is intentionally delivered to that person in order to facilitate change or improvement.

Feedback display – The presentation of scored data, represented graphically or in narrative form.

Graphic display – Charts or graphs which provide a visual numeric portrayal of data.

Halo effect – High correlations resulting from raters' overall bias toward the individual.

Hypothesis – An expectation about observable phenomena derived from theory.

Independent criteria – A criterion variable which comes from a different source or is external to the instrument itself.

Internal consistency – The average correlation among items included in a scale.

Internally homogeneous – Having high internal consistency.

Interrater agreement – Agreement between ratings within rater groups.

Item – A word, phrase or sentence that describes an aspect of behavior, performance, skill, or personal characteristic.

Item content – The behavior, skill, or characteristic to which an item refers.

Item level – Referring to items.

Item-scale correlations – The relationship(s) between responses to individual items and computed scores on scales.

Mean – Arithmetic average of a set of data points or numbers.

Model – A set of hypotheses about the relationships between variables or constructs.

Multimethod study – A study which compares the correlation between two constructs measured by different methods (e.g., rater observation and paper-and-pencil methods). The correlations between the two different

constructs measured similarly should be lower than that between a single construct measured in different ways.

Narrative display – Data presented in text format, usually with personalized interpretation.

Norm group – Group of managers whose scores are stored in the vendor's test database and are output as a comparison on every individual feedback report.

Norms – The average scores of all people who have taken the instrument or the average scores of specific target groups.

Population – The whole universe or group from which samples are drawn.

Psychometrics – The study or practice of numerically estimating aspects of performance or behavior.

Predictive validity – The relationship between scores on the instrument measured at an earlier time and performance or effectiveness in the future.

Rater groups – Groups of individuals providing responses to items on an instrument.

Raters – Individuals providing responses on an instrument.

Rational/intuitive scale development – Grouping items together based on the authors' expectations or experience about how different skills or behaviors relate to one another.

Response scale – A set of choices provided on an instrument for raters to use in evaluating behavior, performance or skill.

Sample – A subset of individuals drawn from the total population or database.

Scale – A group of items that have logical and empirical coherence.

Standard deviation – An index that specifies the amount of spread in the distribution of scores about the mean.

Statistical scale development – Using statistical-clustering procedures such as factor analysis, cluster analysis, or item-scale correlations to group items into scales based on the degree of similarity in the response patterns of raters.

Target population – A group, sometimes with particular characteristics, for whom the instrument is intended; the group that will be rated on an instrument.

Test-retest reliability – Stability of scores over short periods of time.

Validity – Proper performance of an instrument; the property of performing as intended.

Validation strategy – The process or set of studies used to collect evidence regarding the uses to which an instrument can be reasonably and accurately applied.

Variable – Any measurable entity whose value is not constant or given.

Weighting – Assigning differential value to points on a scale.

INSTRUMENT EVALUATION CHECKLIST

GETTING STARTED

(1) Find out what is available.

Have you done a thorough search for management or leadership development instruments? Check the source book(s) you've examined.

____ *Directory of Human Resource Development Instrumentation*

____ *Mental Measurements Yearbook*

____ *Tests: A Comprehensive Reference for Assessment in Psychology, Education and Business*

____ *Psychware Sourcebook*

____ promotional brochures from instrument publishers

____ other sources

(2) Collect a complete set of materials.

Do you have all of the following for each instrument being reviewed?

____ a copy of the instrument

____ a sample feedback report

____ a technical manual

____ supporting materials (sample development guide, list of trainer materials, etc.)

____ purchasing information (price, scoring, certification, etc.)

3) **Compare your intended use to instrument characteristics.** What are the special needs, job demands, etc., of the level of your target managers?

List below any competencies important to your organization or in the program in which the instrument will be used.

_____

_____

_____

_____

_____

_____

_____

_____

_____

Has the instrument been normed on a sample similar to your target managers, in terms of level, gender, ethnicity, etc?

____ Yes ____ No

(4) Examine the feedback scales.

Are the instrument's scales a good fit with your organization's or training program's management or leadership model?

____ Yes ____ No ____ Maybe

Are the instrument's scales a good fit with your target audience (level and role of employees/managers who will get feedback)?

____ Yes ____ No ____ Maybe

CHECKPOINT: ELIMINATE NOW THOSE INSTRUMENTS NOT MEETING YOUR NEEDS IN TERMS OF TARGET AUDIENCE OR DOMAINS (FEEDBACK SCALES).

(5) Familiarize yourself with instrument-development process.

(6) Learn how items and feedback scales were developed.

What is the source of the instrument's items? (more than one may apply)

____ theory

____ research

____ experience

What is the nature of item content? (more than one may apply)

____ behavioral

____ skill-based

____ trait or characteristic-based

What types of statistical methods were used for scale development?

____ factor analysis

____ cluster analysis

____ item-scale correlations

(7) Find out how consistent scores tend to be.

Internal consistency

____ all scales have coefficients between .6 and .8

Interrater agreement

____ all scales have coefficients between .4 and .7

Test-retest

____ all scales have coefficients greater than .4

(8) **Assess basic aspects of validity: Does the instrument measure what it claims to measure?** (Look in technical manual.)

Found at least one study showing instrument measures what it says it measures (e.g., comparing scores on this instrument to hypotheses about scores or to scores on another instrument)?

____ Yes ____ No

Found at least one study showing scores on the instrument are related to a measure of effectiveness?

____ Yes ____ No

Sample used in research studies reflects characteristics of your target audience?

____ Yes ____ No

If instrument will be used internationally:

Has it been translated and back-translated?

____ Yes ____ No

____ Instrument itself

____ Feedback report

____ Development materials

Are international norms available?

____ Yes ____ No

Have reliability and validity been studied internationally?

____ Yes ____ No

CHECK POINT: ELIMINATE NOW THOSE INSTRUMENTS HAVING LITTLE EVIDENCE OF RELIABILITY OR VALIDITY.

CONSIDER POTENTIAL IMPACT

(9) Think about face validity.

Is item wording clear, manager-oriented, and lacking in gender or ethnic bias?

____ Yes ____ No

(10) Examine the response scale.

Does the length meet your needs?

____ Yes ____ No

Is there a nonresponse option (i.e., "Don't know," "Not applicable")?

____ Yes ____ No

(11) Evaluate the feedback display.

What type of display is used in the feedback?

____ Graphic

____ Narrative

Is the display understandable?

____ Clear and understandable (understood almost immediately)

____ Somewhat difficult to interpret (took me two to three attempts to understand)

____ Very hard to understand (still having trouble after three attempts)

(12) Understand how breakout of rater responses is handled.
(More than one category may apply.)

____ Feedback from boss shown separately

____ Feedback from peers shown separately

____ Feedback from subordinates shown separately

____ No breakout of rater data (all of the above categories combined)

(13) Learn what strategies are used to facilitate interpretation of scores.

____ Comparison to norms

____ Highlighting largest self/rater discrepancies

____ Item-level feedback

____ Highlighting high/low scores

____ Comparison to ideal

____ Importance to job or success

____ "Do more/Do less" feedback

CHECK POINT: AT THIS TIME, ELIMINATE THOSE INSTRUMENTS HAVING ITEMS WHICH DO NOT MAKE SENSE OR ARE OFFENSIVE, HAVE RESPONSE SCALES THAT DO NOT MEET YOUR NEEDS (IN TERMS OF LENGTH OR NONRESPONSE OPTIONS), OR HAVE STRATEGIES FOR INTERPRETING SCORES THAT ARE DIFFICULT TO USE OR UNDERSTAND.

(14) Look for development and support materials.

Is there support for the manager's development?

____ Development guide

____ Goal-planning work sheets

____ Workshop available

____ Vendor hotline

Is there trainer support?

____ Trainer's manual

____ Workshop available

____ Overheads

____ Video

____ Supplemental norms available

Is certification required?

____ Yes ____ No

How is scoring done?

____ Manually scored

____ Computer-scored

____ Vendor-scored

____ Computer software

OTHER CONSIDERATIONS

(15) Compare cost: value for the price.

Does the price of the instrument compare favorably to others and is it within the range you can afford?

____ Yes ____ No

(16) Consider length a minor issue.

Is the instrument either excessively long or too brief to be useful and acceptable for your managers?

____ Yes ____ No

OTHER REPORTS IN THE SERIES

Leaders and Leadership: Of Substance and Shadow
Morgan W. McCall, Jr. (1977, Stock #102R) $10.00

Appraisal in Context: Clashing With Organizational Realities
Morgan W. McCall, Jr., & David L. DeVries (1977, Stock #104R) $10.00

Leadership as a Design Problem
Morgan W. McCall, Jr. (1978, Stock #105R) $10.00

Looking at Leadership: Some Neglected Issues
Michael M. Lombardo (1978, Stock #106R) $12.00

Power, Influence, and Authority: The Hazards of Carrying a Sword
Morgan W. McCall, Jr. (1978, Stock #110R) $12.00

Looking Glass, Inc.®: An Organizational Simulation
Morgan W. McCall, Jr., & Michael M. Lombardo (1978, Stock #112R) $7.50

In Pursuit of the Manager's Job: Building on Mintzberg
Morgan W. McCall, Jr., & Cheryl Segrist (1980, Stock #114R) $5.00

Leadership and the Professional
Morgan W. McCall, Jr. (1981, Stock #117R) $10.00

Off the Track: Why and How Successful Executives Get Derailed
Morgan W. McCall, Jr., & Michael M. Lombardo (1983, Stock #121R) $10.00

Trade Routes: The Manager's Network of Relationships
Robert E. Kaplan, & Mignon S. Mazique (1983, Stock #122R) $10.00

A Mirror for Managers: Using Simulation to Develop Management Teams
Robert E. Kaplan, Michael M. Lombardo, & M. S. Mazique (1983, Stock #123R) $12.00

Fixing Relationships Through Joint Action
Robert E. Kaplan (1983, Stock #124R) $10.00

High Hurdles: The Challenge of Executive Self-Development
Robert E. Kaplan, Wilfred H. Drath, & Joan R. Kofodimos (1985, Stock #125R) $15.00

Developmental Experiences in Managerial Work: A Literature Review
Cynthia D. McCauley (1986, Stock #126R) $10.00

The Warp and Woof of the General Manager's Job
Robert E. Kaplan (1986, Stock #127R) $12.00

Values in Action: The Meaning of Executive Vignettes
Michael M. Lombardo (1986, Stock #128R) $12.00

Anatomy of an Executive
Joan R. Kofodimos, Robert E. Kaplan, & Wilfred H. Drath (1986, Stock #129R) $25.00

Creativity in the R&D Laboratory
Teresa M. Amabile, & Stanley S. Gryskiewicz (1987, Stock #130R) $12.00

Explorations of Scope and Scale: The Critical Determinant of High-Level Executive Effectiveness
V. Jon Bentz (1987, Stock #131R) $12.00

Key Events in Executives' Lives
Esther H. Lindsey, Virginia Homes, & Morgan W. McCall, Jr. (1987, Stock #132R) $65.00

Developing Executives Through Work Experiences
Morgan W. McCall, Jr. (1988, Stock #133R) $10.00

The Dynamics of Management Derailment
Michael M. Lombardo, & Cynthia D. McCauley (1988, Stock #134R) $12.00

Eighty-eight Assignments for Development in Place: Enhancing the Developmental Challenge of Existing Jobs
Michael M. Lombardo, & Robert W. Eichinger (1989, Stock #136R) $12.00

Why Executives Lose Their Balance
Joan R. Kofodimos (1989, Stock #137R) $15.00

Preventing Derailment: What To Do Before It's Too Late
Michael M. Lombardo, & Robert W. Eichinger (1989, Stock #138R) $20.00

Using Biographical Methods to Understand Managerial Behavior and Personality
Joan R. Kofodimos (1990, Stock #139R) $15.00

Forging Consensus: Building a Dialogue Among Diverse Leaders
Bernie Ghiselin (1990, Stock #140R) $12.00

Traps and Pitfalls in the Judgment of Executive Potential
Marian N. Ruderman, & Patricia J. Ohlott (1990, Stock #141R) $15.00

Redefining What's Essential to Business Performance: Pathways to Productivity, Quality, and Service
Leonard R. Sayles (1990, Stock #142R) $20.00

Character Shifts: The Challenge of Improving Executive Performance Through Personal Growth
Robert E. Kaplan (1990, Stock #143R) $25.00

Twenty-two Ways to Develop Leadership in Staff Managers
Robert W. Eichinger, & Michael M. Lombardo (1990, Stock #144R) $12.00

Gender Differences in the Development of Managers: How Women Managers Learn From Experience
Ellen Van Velsor, & Martha W. Hughes (1990, Stock #145R) $30.00

Effective School Principals: Competencies for Meeting the Demands of Educational Reform
Cynthia D. McCauley (1990, Stock #146R) $15.00

The Expansive Executive (Second Edition)
Robert E. Kaplan (1991, Stock #147R) $20.00

Understanding Executive Performance: A Life-Story Perspective
Charles J. Palus, William Nasby, & Randolph D. Easton (1991, Stock #148R) $15.00

Feedback to Managers, Volume I: A Guide to Evaluating Multi-rater Feedback Instruments
Ellen Van Velsor, & Jean Brittain Leslie (1991, Stock #149R) $15.00

Feedback to Managers, Volume II: A Review and Comparison of Sixteen Multi-rater Feedback Instruments
Ellen Van Velsor, & Jean Brittain Leslie (1991, Stock #150R) $75.00

Feedback to Managers, Volumes I and II
Ellen Van Velsor, & Jean Brittain Leslie (1991, Stock #151R) $80.00

Motivation in Business and Management: An Information Sourcebook
Karen Rickards Hardie (1990, Stock #315R) $30.00

Readers' Choice: A Decade of *Issues & Observations*
Wilfred H. Drath, Editor (1990, Stock #314R) $15.00

Organization Development: An Annotated Bibliography for the Practitioner
Karen Rickards Hardie, & Ron G. Harrison (1987, Stock #311R) $20.00

The Intuitive Pragmatists: Conversations with Chief Executive Officers
James S. Bruce (1986, Stock #310R) $12.00

The Use of Personality Measures in the Leadership Development Program
David P. Campbell, & Ellen Van Velsor (1985, Stock #308R) $10.00

Coping With an Intolerable Boss
Michael M. Lombardo, & Morgan W. McCall, Jr. (1984, Stock #305R) $10.00

A Selected Bibliography on Employee Attitude Surveys
Margaret C. Blasingame (1981, Stock #302R) $12.00

Discounts are available. Please write for a comprehensive Resource Guide (reports, books, videotapes, and audiotapes). Address your request to: Publication, Center for Creative Leadership, P. O. Box 26300, Greensboro, NC 27438-6300, 919-288-7210, ext. 2805. All prices subject to change.

12/91

ORDER FORM

Name _____ Title _____

Organization _____

Mailing Address _____

City/State/Zip _____

Telephone _____

Quantity	Stock No.	Title	Unit Cost	Amount
			Subtotal	
			Shipping and Handling (Add 5% of subtotal–must be at least $3.00)	
			All NC Residents add 6% sales tax	
			TOTAL	

METHOD OF PAYMENT

☐ Check or money order enclosed (payable to Center for Creative Leadership).

☐ Purchase Order No. _____ (Must be accompanied by this form.)

☐ Charge my order, plus shipping, to my credit card: ☐ VISA ☐ MasterCard ☐ American Express

ACCOUNT NUMBER: _____ EXPIRATION DATE: MO. ____ YR. ____

NAME OF ISSUING BANK: _____

SIGNATURE _____

☐ Please put me on your mailing list.
☐ Please send me the Center's quarterly publication, *Issues & Observations.*

Publications • Center for Creative Leadership • P.O. Box 26300 • Greensboro, NC 27438-6300
919-545-2805 • FAX 919-288-3999

10/91

fold here

PLACE
STAMP
HERE

CENTER FOR CREATIVE LEADERSHIP
PUBLICATIONS
P. O. Box 26300
Greensboro, NC 27438-6300

OUR MISSION

The Center for Creative Leadership is an international, nonprofit educational institution founded in 1970 in Greensboro, North Carolina. Our mission is to encourage and develop creative leadership and effective management for the good of society overall.

We accomplish our mission through research, training, and publication—with emphasis on the widespread, innovative application of the behavioral sciences to the challenges facing the leaders of today and tomorrow.

OUR VALUES

Our work should serve society. We expect our work to make a difference in the quality of leadership in the world. To that end, we try to discover what is most important to do, and focus our resources for the greatest, most enduring benefit. In doing this we continually remind ourselves of the inherent worth of all people. We consider it our responsibility to be attentive to the unique needs of leaders who are women or members of minorities. To make a difference in the world and to turn ideas into action, we must be pioneers in our field, contributors of knowledge, creators of solutions, explorers of ideas, and risk-takers in behalf of society.

Our mission and our clients deserve our best. We expect our service to our clients to be worthy, vigorous, resourceful, courteous, and reliable. In the pursuit of our mission, we intend to be a healthy, creative organization with the financial and inner resources needed to produce our best work. We require ourselves to abide by the highest professional standards and to look beyond the letter of professional guidelines to their spirit. This includes being forthright and candid with every client and program participant, scrupulously guarding the confidentiality of sensitive personal and organizational information, and truthfully representing our capabilities to prospective clients.

Our organization should be a good place to work. To demand the best of ourselves, and to attract, stimulate, and keep the best people, we believe we must make an environment that will support innovation, experimentation, and the taking of appropriate risks. As an organization we should prize the creative participation of each member of our staff. We should welcome the open exchange of ideas and foster the practice of careful listening. We have a duty to actively encourage the personal well-being and the professional development of every person who works here. We should, therefore, maximize the authority and responsibility each person has to continue to make an ever greater contribution. Our policies should be implemented sensitively and consistently.

We should do our work with regard for one another. We recognize the interdependence of everyone who works here, and we expect ourselves to treat one another with respect, candor, kindness, and a sense of the importance of teamwork. We should foster a spirit of service within the staff so that we may better serve the world at large.

The Center for Creative Leadership does not discriminate with respect to the admission of students on the basis of race, sex, color, national or ethnic origin, nor does it discriminate on any such basis with respect to its activities, programs, or policies.

CENTER FOR CREATIVE LEADERSHIP®

Center for Creative Leadership • Post Office Box 26300 • Greensboro, NC 27438-6300